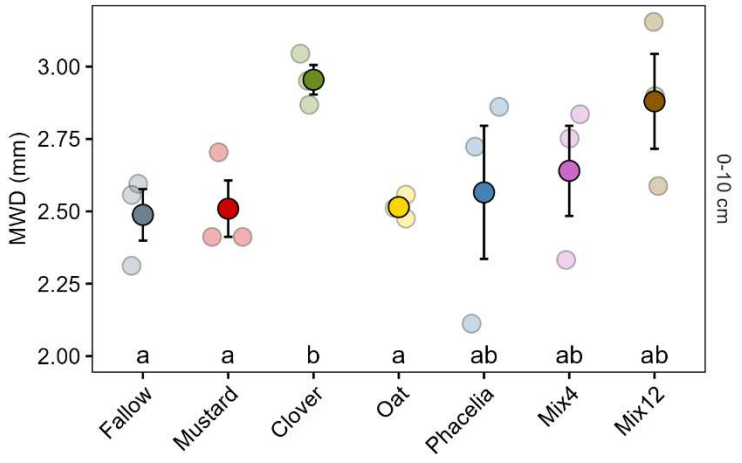
Response to Reviewer #1

We thank the reviewer for the constructive discussion. We took all the comments very seriously and improved the manuscript further as suggested. In the following the comments of the reviewers are numbered and in cursive followed by our response in plain. Changes in the manuscript can be followed in the revision mode. A new R script was compiled and uploaded to the server.

Main comments:

1. *Why was the pairwise t-tests chosen instead of analysis of variance (ANOVA) to compare the significant difference between the seven treatments, e.g., Fig. S5?*

There are, of course, various different methods to evaluate differences between treatments. Pairwise t-tests and ANOVA followed by a post hoc test (usually Tukey's HSD) are the most commonly used and there is no clear rule what to use in which situation. Usually both methods can be used in exchange, but there are some subtle differences. First, F statistic of an ANOVA is not always a robust test. Particularly if the variances across the groups are not equal. Thus, the assumptions that F-statistic is reliable is the assumption that the variances of the groups are equal. Let's focus for example on the following example below (Fig.2a upper facet 0-10 cm):



We see by eye that the variances of the treatments are very different to each other. This is one of the reasons why we show all the data in the plot, not just mean and SE. We can run a Leven's test to prove that:

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  6  0.5667 0.7502
      14
```

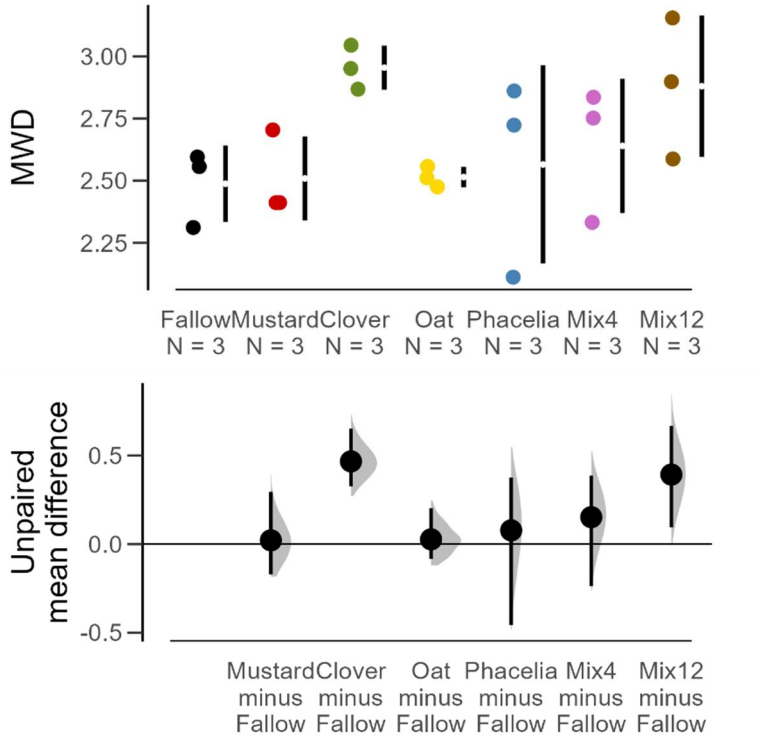
As you can see the assumption of homogeneity of variances is violated because Pr(>F) is larger 0.05, and F value is below 1. So, we cannot run an ANOVA nor a Student's t-Test (same assumption as ANOVA). The only option is a Welch's t-Test which does not have the assumption of homogeneity of variances. This is what we did with the *paiwiseTest* function from the *pairwiseCI* R package.

```
P-values calculated using
 Welch Two Sample t-test

                    p.value
```

```
Mustard-Fallow     0.8797
Clover-Fallow      0.0173
Oat-Fallow         0.7949
Phacelia-Fallow    0.7769
Mix4-Fallow        0.4554
Mix12-Fallow       0.1238
Clover-Mustard     0.0268
Oat-Mustard        0.9607
Phacelia-Mustard   0.8378
Mix4-Mustard       0.5228
Mix12-Mustard      0.1398
Oat-Clover         0.0054
Phacelia-Clover    0.2294
Mix4-Clover        0.1718
Mix12-Clover       0.7002
Phacelia-Oat       0.8463
Mix4-Oat           0.5066
Mix12-Oat          0.1531
Mix4-Phacelia      0.8037
Mix12-Phacelia     0.3341
Mix12-Mix4         0.3482
```

Another option would be a relatively new approach of estimation plots introduced by (Ho et al., 2019). Estimation plots have the advantage not to depend on p-values and null-hypothesis significance testing. According to this method the example from above would look like the following:



According to the estimation plots not only clover would be different from the fallow (as the Welch test above showed) but also Mix12. But we have refrained from using this kind of statistic across all different comparisons. It would be too complex and overloading in the Manuscript.

2. *It is very confusing to understand the results of significant difference between all indexes because the description in main text did not always show the consistency with figures, e.g., Fig. 2, Fig. S4, S5, S6, S7. For example, 'The significantly higher MWD was observed for*

clover at 0-10 cm (18.8% higher), Mix12 at 20-30 cm (37.6% higher) compared to the fallow in Line 200-201 (Fig. 2)'. However, lowercase letters of significant difference are 'b' for clover and 'a' for Mix12 in Fig. 2a.Please check and keep consistency throughout the manuscript.

Thank you for the hint. Indeed, it appeared to be a bit confusing. Note, the pairwise comparison was done for each soil horizon separately. So, it does not matter for the scientific correctness how the order of the levels is. I marked that in the manuscript as well. Nevertheless, I wrote a new function that reordered the levels of the data so that the letters appear in an ascending order of the mean. From now on, the letter "a" starts always with the treatment with the lowest mean. It is, however, not possible to order the levels in a in a specific fixed order of the treatments. This would be possible only manually during plotting. As we wanted to have a machine-readable script, that could be accessed from everyone without explanation, we refrain from a manual sorting. The new function was applied to all figures that used pairwise t-tests. An updated R script was uploaded to the Zenodo server.

Specific comments:

1. It would be better to add 1-2 sentences in Abstract to indicate the OC distribution within macroaggregate fractions.

We added 2 sentences.

2. Please add the details of planting date and harvest date for all crops in the section of Materials and Methods. For CC mixture treatment could add the ranges of planting and harvest date for cover crops.

The information were added as supplementary Table S3.

3. 'OC2_1' is not defined in Line 163-164. Please check and add the missing information.

It is explained in the next sentences: "Principal component analyses confirmed a similar loading of OC2_1 and OC4_2 on the first two components (eigenvalue >1), explaining 61% of the variance in the data. As OC2_1 and OC4_2 are redundant variables, including both does not fit to the model structure. Thus, we excluded OC2_1 from the latent variable construction."

4. Please add the statistical analysis of significant difference and the range of p value in the section of Materials and Methods.

Done.

5. Please add subheadings for each independent results in the section of Results, e.g., 3.1 SOC concentrations and stocks. 3.2 Soil aggregates distribution….

Subheadings were included.

6. Line 215. Please check the figure caption in Fig. S10. 'MWD' change as 'GMD'?

No, everything is correct. Maybe it was a bit confusing because we finished the last paragraph the GMD.

7. The first paragraph in Discussion section is overlap with the introduction, objectives of the study and materials. Please rewrite.

Yes, I always like to recall the basic aims of the study before starting the discussion. But I agree that the text is redundant an removed it.

8. Line 344-350. Please refined and delete the citation. It would be better if the authors could add 1or 2 statements to expand the future research recommendations.
Done.

9. The document of supplementary material S2 is missing. Please check.

Right. S2 is the R markdown file. I wanted to include this file after the review when the script is finalized.

Figures.
1. Fig. 2. Please delete 'Lowercase letters denote' before '(a)… and (b)…' in figure caption.

We changed the description

2. There is no citation for Fig. S11 in the main text.

There was a mistake in line 225. I wrote Fig. S10 instead S11. It is corrected now.

References:

Ho, J., Tumkaya, T., Aryal, S., Choi, H., and Claridge-Chang, A.: Moving beyond P values: data analysis with estimation graphics, Nat Methods, 16, 565–566, https://doi.org/10.1038/s41592-019-0470-3, 2019.